The importance of replicating genomic analyses to verify phylogenetic signal for recently-evolved lineages.

Ceridwen I Fraser[1], Angela McGaughran[2], Aaron Chuah[3], Jonathan M Waters[4]

1. Fenner School of Environment and Society, Australian National University, Canberra, ACT 2601, Australia
2. CSIRO Land and Water, Black Mountain Laboratories, Clunies Ross Street, ACT 2601, Australia; and University of Melbourne, School of BioSciences, 30 Flemington Road, VIC 3010, Australia
3. John Curtin School of Medical Research, Australian National University, Canberra, ACT 2601, Australia
4. Allan Wilson Centre for Molecular Ecology and Evolution, Department of Zoology, University of Otago, Dunedin 9016, New Zealand

Genotyping by Sequencing (GBS); Single Nucleotide Polymorphism (SNP); kelp; macroalgae; marine; speciation

Corresponding author: Ceridwen Fraser

Address: Fenner School of Environment and Society, Australian National University, Canberra, ACT 2601, Australia.

Email: ceridwen.fraser@gmail.com

Fax: +61-2-61250746

Running title: Resolving closely-related species using SNPs

Abstract:

Genome-wide SNP data generated by non-targeted methods such as RAD and GBS are increasingly being used in phylogenetic and phylogeographic analyses. When these methods are used in the absence of a reference genome, however, little is known about the locations and evolution of the SNPs. In using such data to address phylogenetic questions, researchers risk drawing false conclusions, particularly if a representative number of SNPs is not obtained. Here, we empirically test the robustness of phylogenetic inference based on SNP data for closely-related lineages. We conducted a genome-wide analysis of 75,712 SNPs, generated via GBS, of southern bull-kelp (*Durvillaea*). *Durvillaea chathamensis* co-occurs with *D. antarctica* on Chatham Island, but the two species have previously been found to be so genetically similar that the status of the former has been questioned. Our results show that *D. chathamensis*, which differs from *D. antarctica* ecologically as well as morphologically, is indeed a reproductively isolated species. Furthermore, our replicated analyses show that *D. chathamensis* cannot be reliably distinguished phylogenetically from closely-related *D. antarctica* using subsets (ranging in size from 400 to 40,912 sites) of the parsimony-informative SNPs in our dataset, and that bootstrap values alone can give misleading impressions of the strength of phylogenetic inferences. These results highlight the importance of independently replicating SNP analyses to verify that phylogenetic inferences based on non-targeted SNP data are robust. Our study also demonstrates that modern genomic approaches can be used to identify cases of recent or incipient speciation that traditional approaches (e.g., Sanger sequencing of a few loci) may be unable to detect or resolve.

Introduction

High-throughput DNA sequencing technologies are becoming increasingly popular for phylogenetic analysis. For non-model organisms, obtaining large amounts of genomic data for phylogenetic analysis has, however, proven challenging, as tools such as universal primers have only been developed for relatively few phylogenetically-informative regions (Faircloth et al. 2012), and targeted enrichment approaches such as exon capture require knowledge of the genome of the study taxon (Faircloth et al. 2012; Hugall et al. 2016) or close relatives (Bragg et al. 2015). In contrast, SNP data generated by non-targeted methods such as RAD tag (Restriction-site-Associated DNA tags: Miller et al. 2007; Baird et al. 2008) and GBS (Genotyping-by-Sequencing: Elshire et al. 2011) offer appealing alternatives to targeted phylogenetic methods, including for species delimitation (Bryant et al. 2012; Leaché et al. 2014; Herrera and Shank 2015; Pante et al. 2015), as large amounts of data – tens to hundreds of thousands of SNPs – can be obtained with no prior knowledge of the genome. Genomic analyses of closely-related sympatric lineages can, for example, help to detect neonascent but reproductively isolated species, shedding light on the evolutionary processes driving speciation where traditional methods such as sequencing a few, targeted loci might fail (Pante et al. 2015).

When SNPs are obtained by non-targeted methods, and in the absence of a reference genome, little or nothing is known about the genomic locations of the SNPs, precluding inference of their positional-based evolutionary dynamics, and thus potentially violating the assumptions of downstream phylogenetic analyses. Among-site rate variation in a genome can, for example, be substantial: in mitochondrial genomes, some sites have been inferred to have evolved up to 1000 times faster than others (e.g., Galtier et al. 2006; Kjer and Honeycutt 2007; Rosset et al. 2008; Song et al. 2010). Such rate variation can lead to biased estimation of branch lengths, with large impacts on the accuracy of phylogeny estimation, substitution rates, and evolutionary divergence estimates (e.g., Wakeley 1993; Tateno et al. 1994; Yang 1996; Buckley et al. 2001; Sullivan and Swofford 2001; Simon et al. 2006; Soubrier et al. 2012). In datasets generated using random restriction enzyme digests and without a reference genome, SNP sites will correspond to a range of unknown, divergent genomic locations, including coding and non-coding regions; applying appropriate evolutionary models to account for heterogeneity in mutation rates among sites in phylogenetic analyses using *a priori* knowledge of the site characteristics is thus not possible without annotating gene

fragments using genomic data from other species. Such shortcomings could theoretically be offset by using a sufficiently large number of SNPs, but the number of parsimony-informative SNPs – which will be a subset of any dataset – needed to provide an accurate phylogeny, is not clear. In a recent study using shotgun sequencing of gibbon (Hylobatidae) genomes, taxa could be distinguished as effectively using 25,531 SNPs as with random subsets of 200 SNPs (Veeramah et al. 2015), but these analyses were largely looking at deep (intergeneric) divergences which were not always well resolved even using the full dataset; how well small datasets of genome-wide SNPs can resolve relationships of recently-diverged taxa and / or populations remains to be determined.

Geologically recent islands provide ideal natural laboratories for studying speciation processes (Shaw 1996; Mendelson and Shaw 2005). Chatham Island, a small (920 km^2) island situated 650 km east of mainland New Zealand, emerged within the last few million years (Campbell 2008; Heenan et al. 2010), and houses a distinctive biota largely assembled via trans-oceanic dispersal events from mainland New Zealand source populations (Trewick 2000; Goldberg et al. 2008; Heenan et al. 2010; Goldberg and Trewick 2011). The bull-kelp *Durvillaea chathamensis* (Hay 1979a) co-occurs on the island with a widespread congeneric *D. antarctica*. *Durvillaea chathamensis* is non-buoyant and is endemic to Chatham Island, whereas *D. antarctica* is both buoyant and widespread, dominating many rocky shore ecosystems in the Southern Hemisphere, and dispersing long distances via rafting (Fraser et al. 2009; Fraser et al. 2011). The validity of *D. chathamensis* has been questioned based on cladistic analyses (Cheshire et al. 1995), and recent molecular analyses of the genus (Fig. 1) (Fraser et al. 2010) also failed to provide strong evidence in support of its status as a distinct species. Indeed, this latter work revealed a close phylogenetic relationship between *D. chathamensis* and a northern New Zealand clade of *D. antarctica*, including shared alleles at cytoplasmic and nuclear loci (Fig. 1) (Fraser et al. 2010). The two species, which grow side-by-side in the Chatham Island intertidal (Schiel et al. 1995), are nonetheless morphologically distinct (Hay 1979b) and have slightly different ecological niches: *D. antarctica* is hollow-bladed and grows only intertidally, from mid- to low-tide mark, whereas *D. chathamensis* is solid-bladed and mainly occurs sub-tidally, from the low-tide mark to about two metres depth (Hay 1979b). We hypothesized that the two morphotypes represent a case of recent speciation, with isolation of the lineages having occurred too recently to be detectable using traditional Sanger sequencing of standard loci used in phylogenetic analyses (such as cox1,

28S, 18S). Here we use GBS data to: i) assess the genealogical basis for separate recognition of *D. antarctica* and *D. chathamensis*; and ii) assess how many SNPs are needed for our phylogenetic conclusions to be considered robust.

Methods

Sampling

Samples of sympatric *D. antarctica* (n = 23) and *D. chathamensis* (n = 27) were collected from three Chatham Island intertidal localities (Fig. 1, Table 1) at which these taxa grow side by side. In addition, *D. antarctica* samples from five mainland NZ localities (n = 19) — focusing specifically on nearby localities that show particularly close phylogenetic similarity with the Chathams *Durvillaea* assemblage for mtDNA markers (Fraser et al. 2010) (Fig. 1) — and from sub-Antarctic Marion Island (n = 4) and the Falkland Islands (n = 7), were included. Tissue samples were preserved in the field in 96% ethanol, and later dried at 60°C for several hours before being placed in ziplock bags containing silica gel beads.

DNA Extraction

DNA was extracted using the MoBio PowerPlant Pro kit (MoBio, Carlsbad, CA). Brown algal (phaeophycean) tissue can contain polysaccharides that interfere with PCR and DNA digestion, and initial screening of extractions indicated low-purity DNA, so modifications to the extraction protocol were made, as follows. A small (~ 1 mm^2) fragment of dried kelp tissue was softened by soaking in 400 µl dH_2O for two hours at 60°C. Samples were then vortexed in tubes containing steel beads for up to two minutes. PowerPlant PD1, PD2 and RNase A solutions were added according to the manufacturers' instructions. Samples were vortexed briefly, incubated at 65°C for ten minutes, and vortexed again for up to two minutes. 100 µl isopropanol was added to limit precipitation of DNA. Subsequent steps were as per manufacturers' protocols, with final elution in 50 µl PD7 solution. Extracted DNA appeared to still contain some alginates, so samples were further purified using the MoBio PowerClean Pro kit (MoBio, Carlsbad, CA). DNA concentrations were assessed using a Qubit 2.0 Fluorometer and dsDNA High Sensitivity assay (Life Technologies). Each sample yielded a total of 30-50 ng DNA.

SNP Analysis

Genotyping-by-sequencing library preparation followed the protocols of Elshire et al. (Elshire et al. 2011) with modifications. DNA extractions were first dried using a vacuum centrifuge at 45°C, then resuspended in 15 µl dH_2O. To each sample, a uniquely barcoded PstI adapter was added (2.25 ng per sample) (Morris et al. 2011). DNA digestion was performed using 4U PstI-HF (New England Biolabs, Ipswich, MA) (Morris et al. 2011) in 1X CutSmart Buffer™, with incubation at 37°C for two hours. Adapters were ligated with T4 DNA ligase in 1X ligation buffer (New England Biolabs, Ipswich, MA), followed by incubation at 16°C for 90 min and 80°C for 30 min. Purification was performed using a Qiagen MinElute PCR purification kit (Qiagen, Valencia, CA), with elution in 25 µl 1X TE. PCRs were carried out in 50 µl volumes containing 10 µl purified DNA, 1X MyTaq™ HS Master Mix (Bioline), and 1 µM each of PCR primers 5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATC*T and 5'CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCT TCCGATC*T (where * indicates phosphorothioation). PCRs were run in an Eppendorf Mastercycler Nexus under the following conditions: 72°C for 5 min, 95°C for 60 s, and 24 cycles of 95°C for 30 s, 65°C for 30 s, and 72°C for 30s, with a final extension step at 72°C for 5 min. Library concentrations for each sample were assessed using a LabChip GXII (Caliper Life Sciences) and all libraries were pooled (20 ng DNA per sample). Size fractionation of the pooled library was achieved via electrophoresis on a 1.5% agarose gel, with a 300 bp size range from 200 - 500 bp selected for sequencing. Sequencing was carried out on one lane of an Illumina HiSeq 2500.

Reads were assessed for quality and trimmed for Illumina TruSeq2 adaptors using Trimmomatic version 0.32 (Bolger et al. 2014). 112,510,564 paired end reads with identifiable PstI adapters were sequenced. The sequencing protocol (similar to Dussex et al. 2015) employed a combinatorial barcode, which required matching partial-barcodes on both ends of paired-end reads to identify the sample each came from. As the downstream TASSEL version 3.0.167 (Bradbury et al. 2007) pipeline does not work with dual-ended barcodes, unique sample-identifying barcodes (a concatenation of the partial barcodes on the R1 and R2 ends of each read) were replaced on both ends of the reads using custom Python scripts (see Supporting Information) for subsequent processing via TASSEL UNEAK, with default parameters, apart from: restriction enzyme (PstI), minimum number of tags required (5),

error-tolerance rate (0.03), minimum/maximum minor allele frequencies (MAF of 0.05 and 0.5), and minimum/maximum call rates (0 and 1). The UNEAK pipeline does not require a reference genome as it uses network analysis and $\chi 2$ tests to filter matched sequence tags to remove most errors and paralogs (as described in detail by Lu et al. 2013). The pipeline was developed for identifying SNPs from bi-allelic markers, and is thus well suited for use with data from *Durvillaea*, which is diplontic (with a diploid macroscopic stage dominating the life cycle: Thornber 2007).

The SNP dataset resulting from this pipeline had 75,712 sites, with heterozygous positions represented by IUPAC ambiguity codes (e.g., the heterozygous position 'C/T' corresponds to a 'Y' in the alignment). Due to the robustness of the binary distance metric (Jaccard index, Hamers et al. 1989) used in downstream analyses, we were able to utilise all SNPs called by UNEAK without the need to filter any out. Under-represented samples (with fewer than 10,000 reads assigned to them) were removed using custom scripts in R version 3.1.0 (R Development Core Team 2013), leaving 73 samples (Table 1). As is common for GBS datasets (Jarquin et al. 2014), the final filtered genotype matrix had a large amount (96.42%) of missing data. R was also used for principle components analysis (PCoA): see supporting information.

Phylogenetic Analyses

The sites in SNP alignments generated from random restriction enzyme digests, and aligned in the absence of a reference genome, are undoubtedly evolving at a variety of rates (see Introduction), and this variation should be taken into account in any phylogenetic estimation efforts. The most common approach to account for rate heterogeneity in an alignment is to model site-specific rates with a gamma distribution. However, alternative approaches, whereby rates are free to vary without being constrained by a pre-specified distribution, have been shown to out-perform the discrete gamma model (Lartillot and Philippe 2004; Pagel and Meade 2004; Huelsenbeck and Suchard 2007), and these approaches may be particularly well-suited to GBS-based SNP data. The phylogenetic software, IQ-TREE ver. 1.4.1 (Nguyen et al. 2015) has an option that includes the FreeRate model (Soubrier et al. 2012) in its model selection strategy, thus explicitly accounting for rate heterogeneity among sites in a pre-specified distribution-free manner. Thus, we used this software package for all of our phylogenetic analysis.

We first used IQ-TREE to identify the optimal model of evolution for the full dataset of 75,712 SNPs using the -m TESTNEWONLY+ASC option (where the '+ASC' flag is used to account for ascertainment bias in SNP data). IQ-TREE was then executed in full mode to infer phylogenetic trees under the maximum-likelihood (ML) criterion using the identified evolutionary model. This was determined to be GTR+G4+ASC (where G4 refers to a gamma distribution with four rate categories), with the following rate parameters: A-C: 0.875, A-G: 2.823, A-T: 0.809, C-G: 0.537, C-T: 2.895, G-T: 1.000; and base frequencies: A: 0.242, C: 0.253, G: 0.260, T: 0.246; a proportion of invariable sites of: 0.669, and a Gamma alpha shape parameter of: 0.034. Ultra-fast bootstrap approximation (Minh et al. 2013) was used with 10,000 bootstraps to assess node support, and the final tree was evaluated in FigTree ver. 1.4.1 (Rambaut 2009) (Fig. 2). Note that, in line with other phylogenetic software, IQ-TREE assigns the same likelihood to each base of an ambiguous site (including heterozygous positions). This analysis was repeated independently ten times, to assess whether the phylogenetic relationships inferred between *D. antarctica* and *D. chathamensis* were robust, and particularly whether the SNP data supported monophyly for the *D. chathamensis* clade that was indistinguishable from *D. antarctica* in previous analyses of mitochondrial and nuclear loci (Fraser et al. 2010).

As well as including FreeRate models in our analysis to determine the optimal model of evolution, we examined the effects of potential rate heterogeneity among sites by running an additional analysis without the inclusion of a gamma rate distribution (i.e., with an evolutionary model of GTR+ASC compared to GTR+G4+ASC, above). As above, we ran the software in full mode with 10,000 bootstraps. Our aim was to determine the potential effect of rate variation among sites on our final tree topology.

Next, we assessed how consistent the phylogenetic relationships estimated for *D. antarctica* and *D. chathamensis* were, as a function of the number of SNPs used in the analysis. First, we determined in IQ-TREE that the full dataset contained a total of 40,912 parsimony-informative sites. We reduced the dataset to retain only these sites as, although including full sequences can help to improve phylogenetic inference of branch lengths (Leaché et al. 2015), we were primarily interested in tree topology for which only parsimony-informative sites are needed. We used a series of bash commands to randomly extract sites, creating six SNP

datasets of lengths: 400, 1,000, 1,500, 2,000, 4,000, and 10,000, and retaining the original dataset of 40,912 parsimony-informative sites (see Supporting Information for commands). Ten random datasets were generated for each SNP length and each was then run through our IQ-TREE pipeline, as outlined above. From each final phylogenetic tree estimate generated, we extracted the bootstrap value for the node that first connected a *D. chathamensis* individual to a sister *D. antarctica* individual (see Fig. 3). We also assessed whether the overall phylogeny generated was consistent with the results from both the full dataset of 75,712 SNPs, and the full parsimony-informative dataset (40,912 SNPs), generating for each replicate a binary decision (yes / no) concerning the monophyly of *D. chathamensis* as sister to the *D. antarctica* Chatham Island/mainland New Zealand clades (see Results); i.e., whether the reproductive isolation of *D. chathamensis* was supported.

In a final analysis, we assessed the distance between phylogenetic estimates obtained for the various SNP datasets and the full parsimony-informative dataset (i.e., 40,912 informative sites), using the -rf_all function in IQ-TREE to determine the Robinson-Foulds (RF) distance between trees (Robinson and Foulds 1981). This metric measures the distance between unrooted phylogenetic trees according to (A + B), where A is the number of data partitions implied by the first tree but not the second tree, and B is the number of data partitions implied by the second tree but not the first tree. Rather than examining particular clades in isolation, the RF metric takes all tree splits into account. We retrieved the range of RF distances within each set of ten trees for the seven datasets, as well as the range of RF distances between the six reduced-length SNP datasets and the full parsimony-informative dataset. In this way, we determined the phylogenetic error associated with SNP choice; to determine the error associated with phylogenetic construction, we performed a final test, taking one SNP file from each of the ten variously-sized SNP datasets created previously, and running IQ-Tree on that input file over ten replicates. We then calculated the RF distance between all ten trees resulting from each same starting SNP file to assess phylogenetic error for each reduced-length SNP dataset and for the full parsimony-informative dataset, as outlined above.

Introgression

As we are interested in assessing the genealogical basis for separate recognition of *D. antarctica* and *D. chathamensis*, we performed introgression tests using the species delimitation software SNAPP ver. 1.3.0 (Bryant et al. 2012) in BEAST ver. 2 (Bouckaert et

al. 2014). Specifically, we were interested in determining whether we could detect introgression between Chatham Island populations of *D. antarctica* and *D. chathamensis* by generating a species tree for all the samples in our dataset. We generated an input xml file based on a binary file of the 75,712 SNPs (i.e., recoded with 012 coding), and used default settings to run SNAPP. We ran the analysis for 10,000,000 MCMC generations and ensured convergence of the resulting output log file using Tracer ver. 1.6 (http://tree.bio.ed.ac.uk/software/tracer/). We then visualised the posterior distribution of species trees produced, using the DensiTree package associated with BEAST2, and looked for evidence of introgression among taxa.

Results

Phylogenetic analyses based on the full dataset of 75,712 SNPs revealed four closely-related but distinct genotypic assemblages, corresponding to (1) *D. chathamensis*; (2) Chatham Island populations of *D. antarctica*; (3) mainland New Zealand populations of *D. antarctica*, and (4) sub-Antarctic populations of *D. antarctica* (Fig. 2). PCoA analysis also supported these geographic and phylogenetic clusters, with five PCs explaining 84.29% of the total variation (PC1: 26.35%; PC2: 25.36%; PC3: 18.1%; PC4: 8.28%; PC5: 6.2%). Binary PCoA clusters with PC1/PC2 and PC3/PC4 are shown in Fig. 4. Phylogenetic analyses using the full dataset support the distinct phylogenetic status of the two Chatham Island morphotypes, with consistent genome-wide differences between them across multiple sympatric localities. The mainland New Zealand and Chatham Island populations of *D. antarctica* were resolved as monophyletic, with *D. chathamensis* as a sister group within the *D. antarctica* complex (Fig. 2). These analyses also revealed strong spatial genetic differentiation, with distinct geographic localities within each of the major groupings represented by distinct genotypic clusters (Fig. 2), and with phylogeographic partitioning for both species among sites on Chatham Island.

Phylogenetic Uncertainty

Our analysis of rate variation among sites (see Methods) showed there to be no difference – other than minor differences in branch lengths – in topology from phylogenetic analyses with and without the gamma distribution included in our evolutionary model (Fig. 2).

The results of our phylogenetic bootstrapping analyses among reduced-length datasets of only parsimony-informative sites are presented in Fig. 3, where the bootstrap value connecting *D. chathamensis* to its sister clade of *D. antarctica* can be seen for each of our SNP datasets. In Figure 5, the number out of ten replicates for which the topologies returned *D. chathamensis* as a monophyletic group is indicated. If we consider the full parsimony-informative dataset (40,912 sites) to have provided the putatively most accurate phylogenetic estimate (we feel this is reasonable, as the topology matches that from our full analysis of 75,712 sites, where monophyly of D. chathamensis is supported: Fig. 2), then these results can be seen to reveal a large degree of uncertainty in bootstrap support for the node connecting *D. chathamensis* to its sister *D. antarctica* clade for SNP datasets less than 10,000 characters in length (Fig. 3). However, even when bootstrap support at a given node is high, obtaining a tree topology consistent with the full dataset (i.e. returning monophyly of *D. chathamensis*) does not necessarily follow (Fig. 5). For example, although the mean bootstrap support connecting *D. chathamensis* and *D. antarctica* exceeded 90% for datasets of $\geq$2,000 SNPs, the topology only resolved *D. chathamensis* as monophyletic for 6/10, 6/10, and 7/10 replicates for SNP datasets of length 2,000, 4,000, and 10,000, respectively (Figs. 3 and 5). Even within the shorter SNP datasets (e.g., $\leq$ 1,500 SNPs), bootstrap support for the *D. chathamensis* / *D. antarctica* relationship reached as high as 84% in individual replicates when the topology was inconsistent (i.e., *D. chathamensis* was not monophyletic) with the full dataset (both the 75,712 and the 40,912 SNP alignments).

RF distances for the reduced-length datasets reiterate the above findings, reflecting a high degree of topological uncertainty with respect to choice of SNP number. For example, using the full parsimony-informative dataset of 40,912 sites, RF distances (number of topological differences between trees) ranged from 0-14, but the number of partitions disagreeing between the trees generated with replicates of the 10,000 SNP datasets reached as high as 100, and for the 400-SNP datasets, reached as high as 138 (Table 2). As well as within-dataset uncertainty, our analyses identified a high degree of between-dataset uncertainty. For example, the shorter 10,000 SNP datasets resulted in trees that differed by up to 82 partitions from those generated with the full parsimony-informative dataset (Table 2). In our final RF tests examining error with respect to phylogenetic replication, we found that RF distances were higher for phylogenetic trees produced from the same input SNP file when that input file had a smaller number of SNPs. For example, the RF distance ranged from 8-80 for

replicated phylogenetics generated from a single 400-SNP input file, and from 0-14 for a single 40,912-SNP input file (Table 2). As a result, a high degree of topological uncertainty exists with respect to both the number of SNPs utilised in the phylogenetic analysis, and the phylogenetic algorithm itself, although in each case, phylogenetic estimates become more robust / similar as the number of parsimony-informative SNPs increased.

Introgression

Our SNAPP analysis resulted in a species tree that showed no support for introgression between Chatham Island populations of *D. chathamensis* and *D. antarctica* (Fig. 6).

Discussion

Non-targeted SNP data for species delimitation

Our results confirm that SNP data from non-targeted approaches such as GBS have great resolving-potential for phylogenetic analysis, including for the genealogical delimitation of closely-related species. Using the full dataset (75,712 SNPs, of which 40,912 were parsimony-informative), we obtained 100% consistent topology at nodes that separated *a priori* taxonomic (*D. chathamensis vs D. antarctica*) and geographic groupings (for *D. antarctica*: North Island New Zealand, South Island New Zealand, and the sub-Antarctic) (Fig.s 2, 3, 5), implying that these phylogenetic estimates are robust. Under a variety of species concepts (e.g. geneaological; biological; phylogenetic) (Donoghue 1985), our results support the distinct species status of *D. chathamensis*, with strong support for the reciprocal monophyly of *D. chathamensis* and its sister *D. antarctica* clade. Alongside our SNAPP analysis, a lack of genetic intermediates argues against the possibility that their mtDNA and chloroplast sequence similarity might reflect introgression, although evidence of introgression from low levels of recent gene flow could be restricted to specific parts of the genome, and older geneflow might be undetectable in our analyses. Detection of introgression could also be limited by the inability of this approach to distinguish between hemizygotes, where only one allele is sequenced for a particular SNP and individual, and homozygotes (Davey et al. 2013). Further tests for introgression (Twyford and Ennos 2012; Eaton et al. 2015) could be performed in future studies using a greater number of samples from a greater number of populations and *Durvillaea* lineages. Alternatively, incomplete lineage sorting could explain the shared alleles at cytoplasmic and nuclear loci. Genome-wide SNP data have nonetheless allowed us to confirm the species status of *D. chathamensis*, when

previous multilocus DNA (Fraser et al. 2010), and morphological and ecological phylogenetic analyses (Cheshire et al. 1995), had failed to clearly resolve them. These findings highlight the utility of GBS data for resolving phylogenetic relationships among closely-related species, and for detecting recent speciation events. Furthermore, the strong divergences detected among these and other *Durvillaea* lineages in our analyses (*D. antarctica* from mainland New Zealand, and from the sub-Antarctic) support previous suggestions that *D. antarctica* may comprise several as-yet unrecognised species (Fraser et al. 2010).

Our results emphasize, however, the need to be cautious when analysing SNP data and interpreting the resultant phylogenies. Results varied drastically depending on the number of SNPs included in our reduced-length, parsimony-informative site analyses (Fig.s 3 and 5), and these conflicting topologies often received high bootstrap support (Fig. 3). Substantial phylogenetic conflict between replicate ML analyses was particularly apparent for datasets with relatively few SNPs. Indeed, when analyses included 10,000 SNPs or fewer, *D. chathamensis* was often resolved as paraphyletic with respect to its sister taxon *D. antarctica* (NZ / Chatham).

GBS-generated SNP data are inherently patchy, with low coverage and high proportions of missing data (Lu et al. 2013), and this kelp dataset was no exception, with the percentage of missing data in the full dataset ranging from 90-99%, and from 71-99% in the reduced-length datasets. Indeed, the process of removing problematic polysaccharides such as alginates during kelp DNA extractions resulted in the amount of DNA used in our GBS library preparation being low (30 - 50 ng per sample, compared to the 100 ng used by Elshire et al. 2011), which probably affected the number of reads obtained. Restriction enzyme digestion can also vary in effectiveness due to factors such as base-composition heterogeneity among taxa (Scaglione et al. 2012), influencing how much SNP data can be obtained via GBS or RAD tag and making interspecific comparisons particularly prone to having larger amounts of missing data. The coverage and depth of datasets will therefore vary for different taxa, as will the number of SNPs needed to resolve phylogenies. A dataset with a smaller proportion of missing data might be less likely to yield differing phylogenetic topologies when different numbers of SNPs are included. Nonetheless, these results highlight the importance of using as many SNPs as possible, and – importantly – of independently replicating phylogenetic

analyses to assess the robustness of the topology, rather than relying on bootstrap values alone.

SNP data as a tool to detect incipient speciation events: the case of Durvillaea

Several evolutionary studies have indicated that repeated ecologically-driven transitions (Rundle and Nosil 2005; Soria-Carrasco et al 2014) can generate rapid genetic divergence, leading to multiple speciation events over short timeframes. Modelling studies have suggested that reproductive isolation can potentially evolve within fewer than one hundred generations (Hendry et al. 2007). In the case of *Durvillaea*, transitions from hollow-bladed (buoyant) to solid-bladed (non-buoyant) morphology may be an important process driving repeated and ongoing diversification. Other distinctive solid-bladed populations of *D. antarctica* have been recorded at several localities across the Southern Hemisphere range of this taxon, including South America (Ramírez and Santelices 1991) and the sub-antarctic islands, for example Macquarie Island (Klemm and Hallam 1988), Marion Island and Gough Island (Hay 1994). As in the case of *D. chathamensis* and *D. antarctica* on Chatham Island, analysis of mtDNA from solid-bladed morphotypes from Gough, Marion and the Falkland Islands has not shown any notable genetic differences between these and sympatric buoyant plants (Fraser et al. 2010). Although a monophyletic origin for solid-bladed forms of *Durvillaea* was originally proposed (Hay 1979a), both molecular (Fraser et al. 2010) and morphological / ecological (Cheshire et al. 1995) cladistic analyses indicate that solid forms have arisen multiple times in the genus. Genome-wide SNP data represent an ideal tool with which to assess whether the multiple solid-bladed forms present in *D. antarctica* also represent examples of incipient reproductive speciation. Broadly, it seems that parallel divergence underpinned by repeated directional selection (e.g. Albertson et al. 2003; Protas et al. 2006; Soria-Carrasco et al. 2014) represents a key force in driving predictable patterns of biotic evolution.

Sympatric speciation or repeated island invasions?

Our analyses suggest that the Chatham Island lineages of *D. antarctica* and *D. chathamensis* are not each other's closest relatives; instead, *D. antarctica* from Chatham Island and *D. antarctica* from nearby mainland New Zealand appear to be monophyletic, with *D. chathamensis* as a sister group within the broader *D. antarctica* complex (Fig. 2). Double invasion (McPhail 1984) of Chatham Island by *Durvillaea* (rather than sympatric speciation)

thus seems the most likely explanation for this phylogenetic pattern. We propose that oceanic dispersal followed by rapid ecomorphological divergence may explain the rapid evolution of these sympatric congeners, emphasizing the likely role of dispersal and founder speciation in driving diversification. Founder-blocking priority effects could explain the maintenance of phylogeographic structure in highly-dispersive species such as *D. antarctica* (Fraser et al. 2009; Waters et al. 2013). Specifically, despite the vast numbers of *D. antarctica* plants drifting at sea (Garden *et al.* 2014; Smith 2002), colonization events appear most likely to occur when dispersing individuals reach shores that are unoccupied by conspecifics (Fraser et al. 2009; Waters et al. 2013). These data also add to the wealth of evidence for recent colonization of the Chatham Islands from mainland source populations, followed by founder speciation (Trewick 2000; Paterson et al. 2006; Goldberg et al. 2008; Heenan et al. 2010; Goldberg and Trewick 2011).

Acknowledgements

Cameron Hay provided extensive discussions on *Durvillaea* biology. Chatham Island samples were collected by Rebecca Cumming and Raisa Nikula during an expedition funded by a Marsden Grant to JMW. Niccy Aitken and Laura Wilson provided laboratory assistance. Sequencing was carried out by the ACRF Biomolecular Resource Facility at the Australian National University and preliminary bioinformatic processing was performed by Cameron Jack at its Genome Discovery Unit. Lars Jermiin provided valuable discussion regarding the phylogenetic analysis of SNP data. Photographs used in Fig. 1 taken by Cameron Hay (*D. chathamensis*) and CIF (*D. antarctica*). This research was supported by an Australian Research Council Discovery Early Career Research Award (DE140101715 to CIF) and University of Otago Performance Based Research Funding (to JMW).

466 **References**

467 Albertson, R. C., J. T. Streelman, and T. D. Kocher. 2003. Directional selection has shaped
468 the oral jaws of Lake Malawi cichlid fishes. Proc. Natl. Acad. Sci. U. S. A. 100:5252-
469 5257.

470 Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U.
471 Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and genetic
472 mapping using sequenced RAD markers. PLoS ONE 3:e3376.

473 Bolger, A. M., Lohse, M., & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina
474 sequence data. Bioinformatics 30:2114-2120.

475 Bouckaert, R.R., Heled, J., Kuehnert, D., Vaughan, T.G., Wu, C.-H., Xie, D., Suchard, M.A.,
476 Rambaut. A. & Drummond, A.J. 2015. BEAST 2: A software platform for Bayesian
477 evolutionary analysis. PLoS Computational Biology 10: e1003537.

478 Bragg, J. G., S. Potter, K. Bi, and C. Moritz. 2015. Exon capture phylogenomics: efficacy
479 across scales of divergence. Molecular Ecology Resources:n/a-n/a.

480 Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012.
481 Inferring species trees directly from biallelic genetic markers: bypassing gene trees in
482 a full coalescent analysis. Mol. Biol. Evol. 29:1917-1932.

483 Buckley, T. R., C. Simon, and G. K. Chambers. 2001. Exploring among-site rate variation
484 models in a Maximum Likelihood framework using empirical data: effects of model
485 assumptions on estimates of topology, branch lengths, and bootstrap support. Syst.
486 Biol. 50:67-86.

487 Campbell, H. J. 2008. Geology. Pp. 35-52 *in* C. Miskelly, ed. Chatham Islands: heritage and
488 conservation, 2nd Edition. Department of Conservation, Wellington.

489 Cheshire, A. C., J. G. Conran, and N. D. Hallam. 1995. A cladistic analysis of the evolution
490 and biogeography of *Durvillaea* (Phaeophyta). J. Phycol. 31:644-655.

491 Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2013.
492 Special features of RAD sequencing data: implications for genotyping. Mol. Ecol.
493 22:3151–3164.

494 Donoghue, M. J. 1985. A critique of the biological species concept and recommendations for
495 a phylogenetic alternative. The Bryologist 88:172-181.

496 Dussex, N., Chuah, A., & Waters, J.M. 2015. Genome-wide SNPs reveal fine-scale
497 differentiation among wingless alpine stonefly populations and introgression between
498 winged and wingless forms. Evolution 70:38-47.

499 Eaton, D. A. R., A. L. Hipp, A. González-Rodríguez, and J. Cavender-Bares. 2015. Historical
500 introgression among the American live oaks and the comparative nature of tests for
501 introgression. Evolution 69:2587-2601.

502 Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E.
503 Mitchell. 2011. A robust, simple Genotyping-by-Sequencing (GBS) approach for high
504 diversity species. PLoS ONE 6:e19379.

505 Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T.
506 C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers
507 spanning multiple evolutionary timescales. Syst. Biol. 61:717-726.

508 Fraser, C. I., R. Nikula, H. G. Spencer, and J. M. Waters. 2009. Kelp genes reveal effects of
509 subantarctic sea ice during the Last Glacial Maximum. Proc. Natl. Acad. Sci. U. S. A.
510 106:3249-3253.

511 Fraser, C. I., R. Nikula, and J. M. Waters. 2011. Oceanic rafting by a coastal community.
512 Proc. R. Soc. Biol. Sci. Ser. B 278:649-655.

513 Fraser, C. I., D. J. Winter, H. G. Spencer, and J. M. Waters. 2010. Multigene phylogeny of
514 the southern bull-kelp genus *Durvillaea* (Phaeophyceae: Fucales). Mol. Phylogenet.
515 Evol. 57:1301-1311.

516 Galtier, N., D. Enard, Y. Radondy, E. Bazin, and K. Belkhir. 2006. Mutation hot spots in
517 mammalian mitochondrial DNA. Genome Res. 16:215-222.

518 Goldberg, J. and S. A. Trewick. 2011. Exploring phylogeographic congruence in a
519 continental island system. Insects 2:369-399.

520 Goldberg, J., S. A. Trewick, and A. M. Paterson. 2008. Evolution of New Zealand's terrestrial
521 fauna: a review of molecular evidence. Philos. Trans. R. Soc. Lond. B Biol. Sci.
522 363:3319-3334.

523 Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., & Rousseau, R. 1989.
524 Similarity measures in scientometric research: The Jaccard index versus Salton's
525 cosine formula. Inform. Process. Manag. 25:315-318.

526 Hay, C. 1979a. A phytogeographical account of the southern bull kelp seaweeds *Durvillaea*
527 spp. Bory 1826 (Durvilleales Petrov 1965). Proceedings of the International
528 Symposium of Marine Biogeography and Evolution in the Southern Hemisphere,
529 Auckland, New Zealand 2:443-454.

530 Hay, C. 1994. *Durvillaea* (Bory). Pp. 353-384 *in* I. Akatsuka, ed. Biology of Economic
531 Algae. SPB Academic Publishing, The Hague.

532 Hay, C. H. 1979b. Nomenclature and taxonomy within the genus *Durvillaea* Bory
533 (Phaeophyceae, Durvilleales Petrov). Phycologia 18:191-202.

534 Heenan, P., A. Mitchell, P. de Lange, J. Keeling, and A. Paterson. 2010. Late-Cenozoic
535 origin and diversification of Chatham Islands endemic plant species revealed by
536 analyses of DNA sequence data. N. Z. J. Bot. 48:83 - 136.

537 Hendry, A. P., P. Nosil, and L. H. Rieseberg. 2007. The speed of ecological speciation.
538 Funct. Ecol. 21:455-464.

539 Herrera, S. and T. M. Shank. 2015. RAD sequencing enables unprecedented phylogenetic
540 resolution and objective species delimitation in recalcitrant divergent taxa. bioRxiv.

541 Huelsenbeck, J. P. and M. A. Suchard. 2007. A nonparametric method for accommodating
542 and testing across-site rate variation. Syst. Biol. 56:975-987.

543 Hugall, A. F., T. D. O'Hara, S. Hunjan, R. Nilsen, and A. Moussalli. 2016. An exon-capture
544 system for the entire class Ophiuroidea. Mol. Biol. Evol. 33:281-294.

545 Jarquín D, Kocak K, Posadas L, et al. (2014) Genotyping by sequencing for genomic
546 prediction in a soybean breeding population. BMC Genomics 15, 1-10.

547 Kjer, K. M. and R. L. Honeycutt. 2007. Site specific rates of mitochondrial genomes and the
548 phylogeny of eutheria. BMC Evol. Biol. 7.

549 Klemm, M. F. and N. D. Hallam. 1988. Conceptacle development, gamete maturation and
550 embryology of *Durvillaea antarctica* from Macquarie Island. Pap. Proc. R. Soc.
551 Tasman. 122:199-210.

552 Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities
553 in the amino-acid replacement process. Mol. Biol. Evol. 21:1095-1109.

554 Leaché, A. D., B. L. Banbury, J. Felsenstein, A. Nieto-Montes de Oca, and A. Stamatakis.
555 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for
556 inferring SNP phylogenies. Syst. Biol.

557 Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014. Species delimitation
558 using genome-wide SNP data. Syst. Biol. 63:534-542.

559 Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney , M. D. Casler, E. S. Buckler, and
560 D. E. Costich. 2013. Switchgrass genomic diversity, ploidy, and evolution: novel
561 insights from a network-based SNP discovery protocol. PLoS Genetics 9:e1003215.

562 McPhail, J. D. 1984. Ecology and evolution of sympatric sticklebacks (*Gasterosteus*):
563 morphological and genetic evidence for a species pair in Enos Lake, British
564 Columbia. Can. J. Zool. 62:1402-1408.

565 Mendelson, T. C. and K. L. Shaw. 2005. Sexual behaviour: Rapid speciation in an arthropod.
566 Nature 433:375-376.

567 Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and
568 cost-effective polymorphism identification and genotyping using restriction site
569 associated DNA (RAD) markers. Genome Res. 17:240-248.

570 Minh, B. Q., M. A. T. Nguyen, and A. von Haeseler. 2013. Ultrafast approximation for
571 phylogenetic bootstrap. Mol. Biol. Evol.

572 Morris, G. P., P. P. Grabowski, and J. O. Borevitz. 2011. Genomic diversity in switchgrass
573 (*Panicum virgatum*): from the continental scale to a dune landscape. Mol. Ecol.
574 20:4938-4952.

575 Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. IQ-TREE: A fast and
576 effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. Mol.
577 Biol. Evol. 32:268-274.

578 Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-
579 heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571-581.

580 Pante, E., J. Abdelkrim, A. Viricel, D. Gey, S. C. France, M. C. Boisselier, and S. Samadi.
581 2015. Use of RAD sequencing for delimiting species. Heredity 114:450-459.

582 Paterson, A., S. Trewick, K. Armstrong, J. Goldberg, and A. Mitchell. 2006. Recent and
583 emergent: molecular analysis of the biota supports a young Chatham Islands. Pp. 27–
584 29 *in* S. A. Trewick, and M. J. Phillips, eds. Geology and genes III Geological Society
585 of New Zealand, Wellington.

586 Protas, M. E., C. Hersey, D. Kochanek, Y. Zhou, H. Wilkens, W. R. Jeffery, L. I. Zon, R.
587 Borowsky, and C. J. Tabin. 2006. Genetic analysis of cavefish reveals molecular
588 convergence in the evolution of albinism. Nat. Genet. 38:107-111.

589 Rambaut, A. 2009. FigTree.

590 Ramírez, M. E. and B. Santelices. 1991. Catálogo de las algas marinas bentónicas de la costa
591 temperada del Pacífico de Sudamérica. Monografías Biológicas 5:1-437.

592 Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci.
593 53:131-147.

594 Rosset, S., R. S. Wells, D. F. Soria-Hernanz, C. Tyler-Smith, A. K. Royyuru, D. M. Behar,
595 and a. T. G. Consortium. 2008. Maximum-Likelihood estimation of site-specific
596 mutation rates in human mitochondrial DNA from partial phylogenetic classification.
597 Genetics 180:1511-1524.

598 Rundle, H. D. and P. Nosil. 2005. Ecological speciation. Ecol. Lett. 8:336-352.

599 Scaglione, D., A. Acquadro, E. Portis, M. Tirone, S. J. Knapp, and S. Lanteri. 2012. RAD tag
600 sequencing as a source of SNP markers in *Cynara cardunculus* L. BMC Genomics
601 13.

602 Schiel, D. R., N. L. Andrew, and M. S. Foster. 1995. The structure of subtidal algal and
603 invertebrate assemblages at the Chatham Islands, New Zealand. Mar Biol 123:355-
604 367.

605 Shaw, K. L. 1996. Sequential radiations and patterns of speciation in the Hawaiian cricket
606 genus *Laupala* inferred from DNA sequences. Evolution 50:237-255.

607 Simon, C., T. R. Buckley, F. Frati, J. B. Stewart, and A. T. Beckenbach. 2006. Incorporating
608 molecular evolution into phylogenetic analysis, and a new compilation of conserved
609 Polymerase Chain Reaction primers for animal mitochondrial DNA. Annu. Rev. Ecol.
610 Evol. Syst. 37:545-579.

611 Song, H., N. C. Sheffield, S. L. Cameron, K. B. Miller, and M. F. Whiting. 2010. When
612 phylogenetic assumptions are violated: base compositional heterogeneity and among-
613 site rate variation in beetle mitochondrial phylogenomics. Syst. Entomol. 35:429-448.

614 Soria-Carrasco, V., Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman, J. S.
615 Johnston, C. A. Buerkle, J. L. Feder, J. Bast, T. Schwander, S. P. Egan, B. J. Crespi,
616 and P. Nosil. 2014. Stick insect genomes reveal natural selection's role in parallel
617 speciation. Science 344:738-742.

618 Soubrier, J., M. Steel, M. S. Y. Lee, C. Der Sarkissian, S. Guindon, S. Y. W. Ho, and A.
619 Cooper. 2012. The influence of rate heterogeneity among sites on the time
620 dependence of molecular rates. Mol. Biol. Evol. 29:3345-3358.

621 Sullivan, J. and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic
622 inference when we know that assumptions about among-site rate variation and
623 nucleotide substitution pattern are violated? Syst. Biol. 50:723-729.

624 Tateno, Y., N. Takezaki, and M. Nei. 1994. Relative efficiencies of the maximum-likelihood,
625 neighbor-joining, and maximum-parsimony methods when substitution rate varies
626 with site. Mol. Biol. Evol. 11:261-277.

627 Thornber CS (2007) Algal life cycles. In: Encyclopedia of Tidepools and Rocky Shores (eds.
628 Denny MW, Gaines SD), pp. 45-47. University of California Press, Berkeley.

629 Trewick, S. A. 2000. Molecular evidence for dispersal rather than vicariance as the origin of
630 flightless insect species on the Chatham Islands, New Zealand. J. Biogeogr. 27:1189-
631 1200.

632 Twyford, A. D. and R. A. Ennos. 2012. Next-generation hybridization and introgression.
633 Heredity 108:179-189.

634 Veeramah, K. R., A. E. Woerner, L. Johnstone, I. Gut, M. Gut, T. Marques-Bonet, L.
635 Carbone, J. D. Wall, and M. F. Hammer. 2015. Examining phylogenetic relationships
636 among gibbon genera using whole genome sequence data using an approximate
637 Bayesian computation approach. Genetics 200:295-308.

638 Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human
639 mitochondrial DNA. J. Mol. Evol. 37:613-623.

640 Waters, J. M., C. I. Fraser, and G. M. Hewitt. 2013. Founder takes all: density-dependent
641 processes structure biodiversity. Trends Ecol. Evol. 28:78-85.

642 Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends
643 Ecol. Evol. 11:367-372.

644

645 **Data accessibility**: the SNP data set and dataset generation commands are provided as
646 supporting information.

647

648 **Author contributions:** CIF, AM and JMW designed the study; CIF did the laboratory
649 analysis; CIF wrote the first draft; AM and AC analyzed the data. All authors had input on
650 writing the paper and gave final approval for publication.
651

TABLES

Table 1: Number of *Durvillaea* samples per site used in downstream phylogenetic analyses (total: 73)

Clade	**Site / region**	**# Samples**
D. antarctica sub-Antarctic	Falkland Islands	7
	Marion Island	3
D. antarctica New Zealand mainland	Banks Peninsula	5
	Raramai Tunnels	1
	Cape Campbell	4
	Wellington	4
	Maori Bay	2
D. antarctica Chatham Island	Wharekauri	5
	Whangamoe Inlet	8
	Waitangi West	9
D. chathamensis Chatham Island	Wharekauri	9
	Whangamoe Inlet	7
	Waitangi West	9

Table 2: A measure of the range of Robinson-Foulds distances among groups of ten trees estimated for random subsets of parsimony-informative SNPs (SNP choice (within), among the ten trees estimated for each listed SNP subset and the ten trees generated with the full parsimony-informative (40,912 sites) dataset (SNP choice (across); and among sets of ten trees derived from the same starting input SNP file (phylogenetic error).

No. of SNPs in dataset	**SNP choice (within)**	**SNP choice (across)**	**Phylogenetic error**
400	118-138	110-130	8-80
1,000	106-132	96-122	0-36
1,500	88-124	74-114	0-40
2,000	96-124	74-104	0-58
4,000	80-110	66-94	0-38
10,000	64-100	56-82	0-58
40,912	0-14	-	0-14

Figure legends

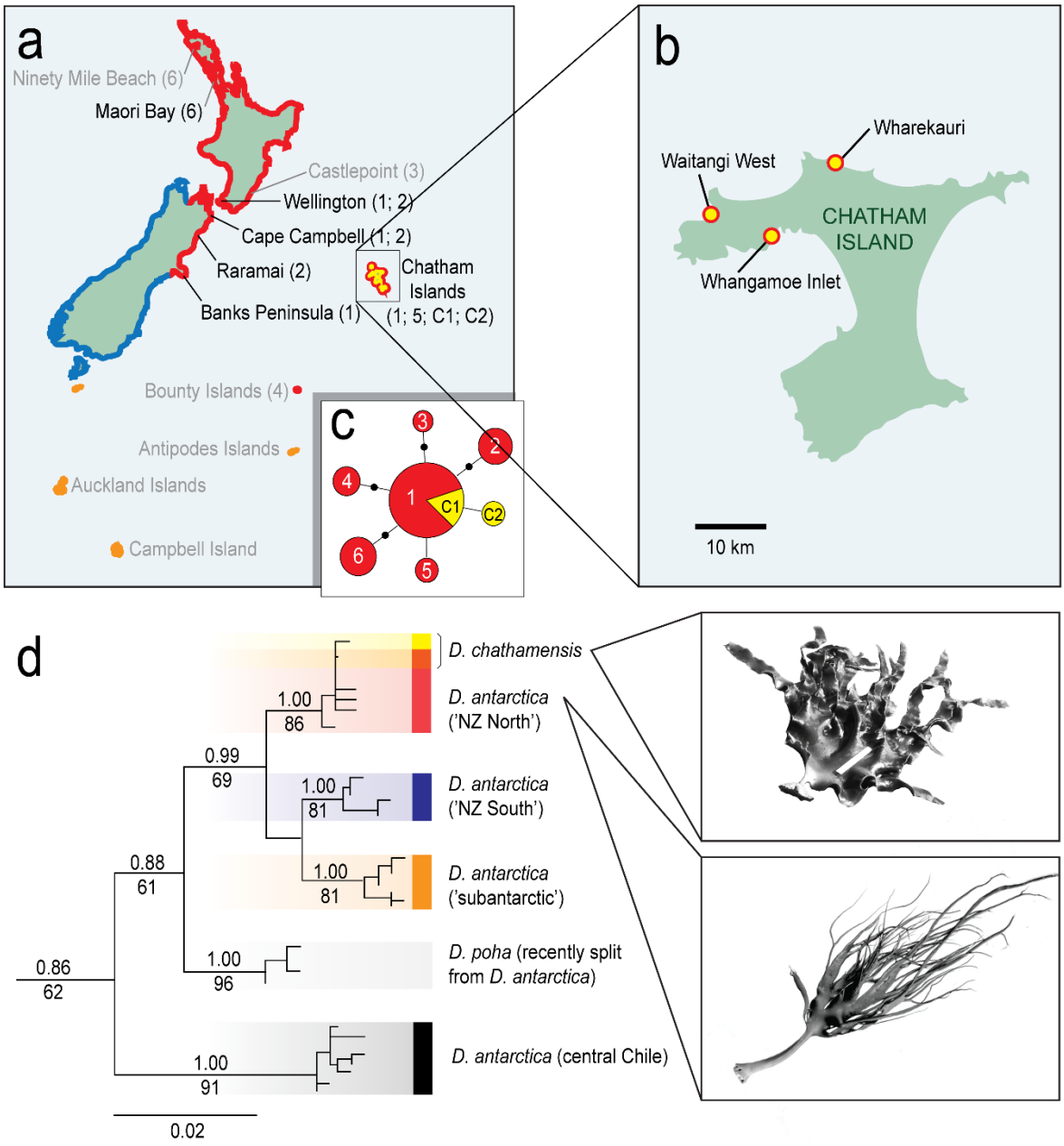

Figure 1: Distribution, phylogeny, and sampling of *Durvillaea antarctica* and *D. chathamensis.* a) distributions of lineages, with colors corresponding to those in other panels. cox1 haplotypes for each site sampled from the *D. chathamensis* / *D. antarctica* 'NZ North' clade (Fraser et al. 2010) are indicated in parentheses after site names, and sites from which samples were used in this study are shown in black text. b) Sampling sites for sympatric *D. antarctica* and *D. chathamensis* used in this study. c) mtDNA (cox1) haplotype network for the *D. chathamensis* / *D. antarctica* 'NZ North' clade (for all samples used in Fraser et al. 2010). d) mtDNA (cox1) phylogeny of the *D. antarctica* / *D. chathamensis* / *D. poha* clade

(from Fraser et al. 2010). Photographs illustrate the morphological differences between *D. antarctica* and *D. chathamensis*.

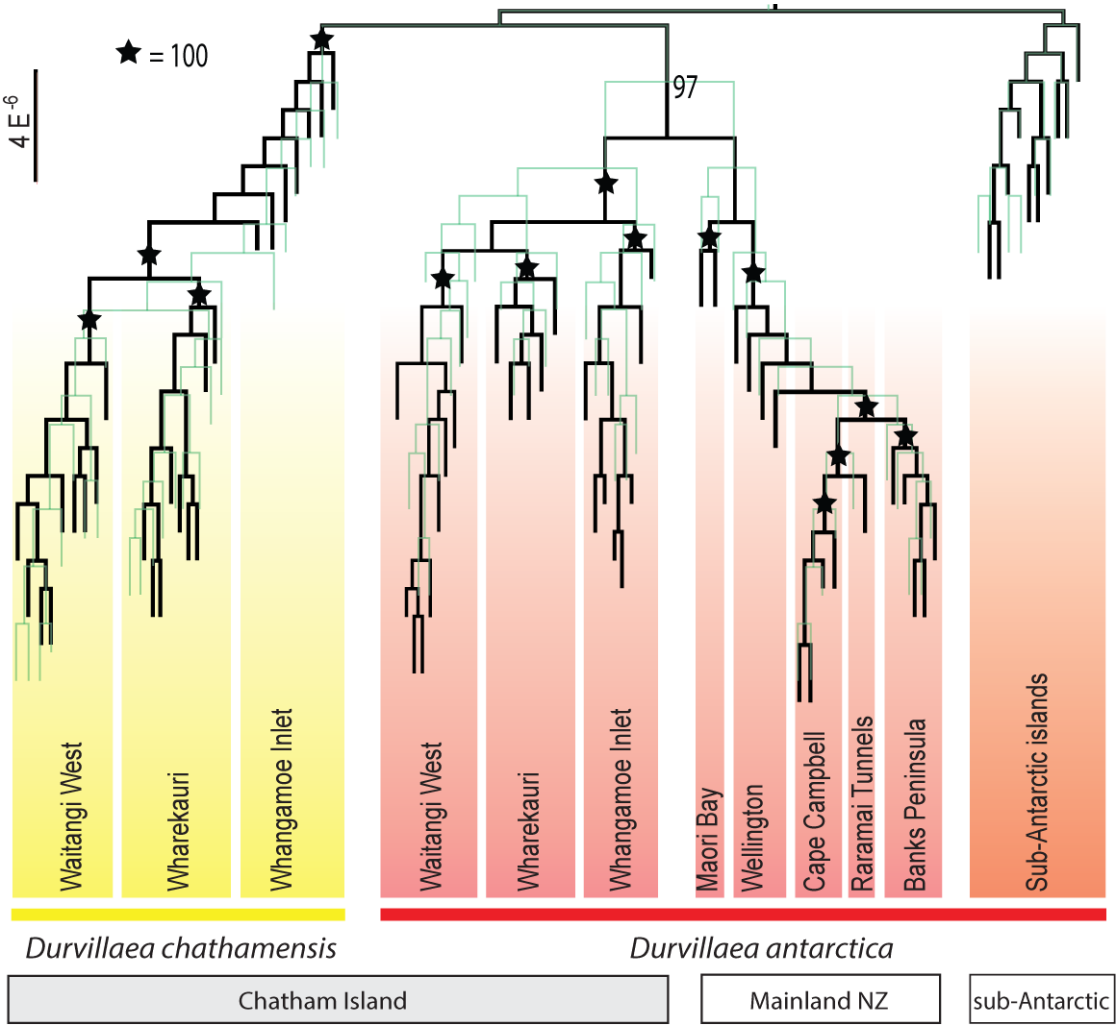

Figure 2: Maximum Likelihood tree for Chatham Island and mainland NZ *Durvillaea* populations based on the full dataset of 75,712 SNPs (black lines). Node support (10,000 bootstrap replicates) is shown for major branches. ML tree for a second analysis using an evolutionary model without gamma rate variation is shown by underlying thin green lines.

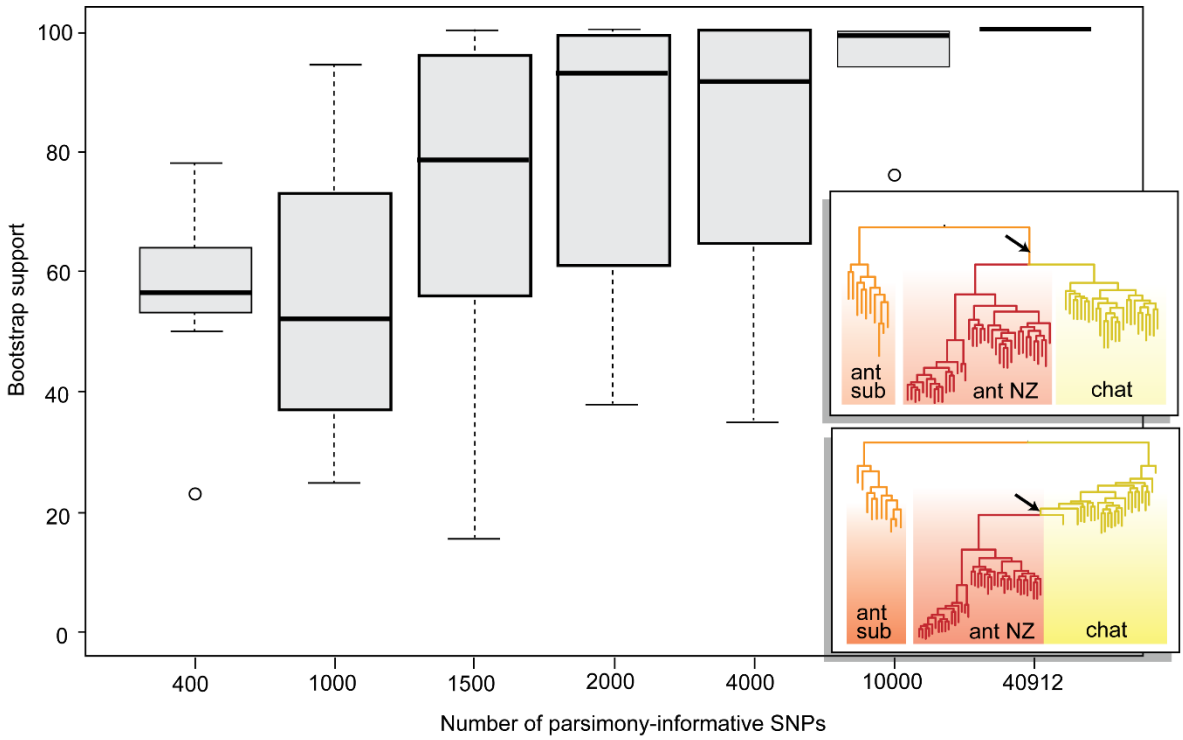

Figure 3: Bootstrap support as a function of number of SNPs used in analyses. Boxplots indicate the range of bootstrap support at the node connecting *D. chathamensis* to its sister clade (*D. antarctica* NZ/Chatham) for ten independent replicate analyses using randomly selected, parsimony-informative subsets of the data (400, 1,000, 1,500, 2,000, 4,000, 10,000 SNPs) and the full parsimony-informative dataset (40,912 SNPs). Inset: examples to demonstrate the location of the node (marked with an arrow) connecting *D. chathamensis* ('chat') to its sister *D. antarctica* clade ('ant NZ,' from the New Zealand mainland and Chatham Island; the sub-Antarctic clade is labelled 'ant sub') in the case of a phylogenetic tree where the monophyly of *D. chathamensis* is supported (upper inset) and not supported (lower inset).

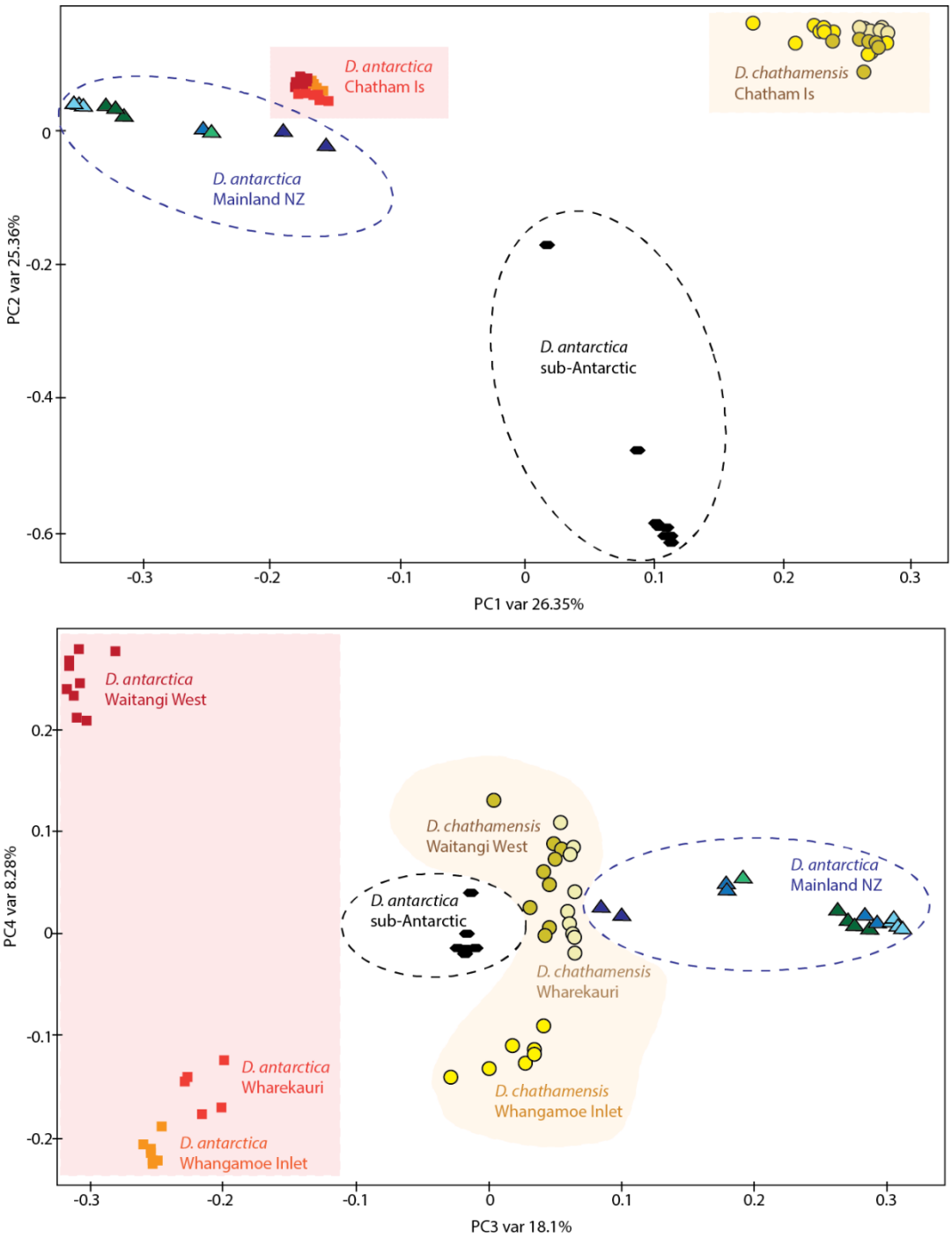

Figure 4: PCoA plots showing geographic and phylogenetic clusters. The regions occupied by the two Chatham Island groups (*D. antarctica* and *D. chathamensis*) are indicated by red and orange shading, with individuals shown as square and circular symbols, respectively. The regions occupied by the two outgroups, *D. antarctica* from the New Zealand mainland and from the sub-Antarctic, are circled, with individuals indicated by triangular and hexagonal symbols, respectively.

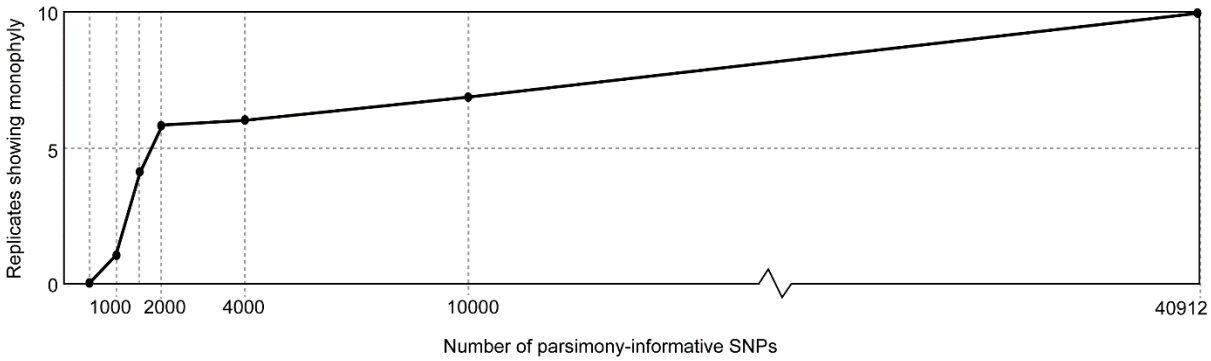

Figure 5: The number out of each set of ten replicates which displayed a final tree topology where *D. chathamensis* was recovered as a monophyletic group, sister to the Chatham and New Zealand mainland *D. antarctica* clades, consistent with the topology resolved from the full dataset (i.e. both the 75,712 full, and the 40,912 parsimony-informative, alignments).

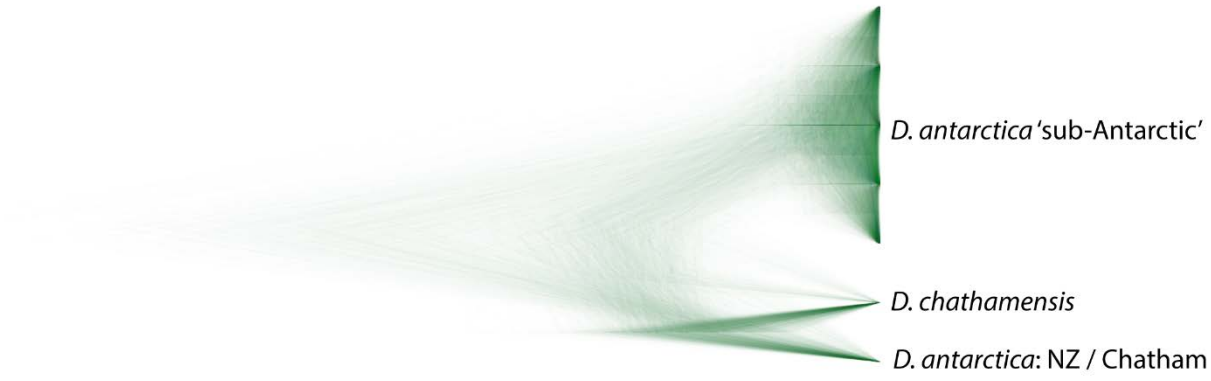

Figure 6: Species delimitation analysis (DensiTree), showing no evidence for introgression between *D. chathamensis* and Chatham Island *D. antarctica* clades.